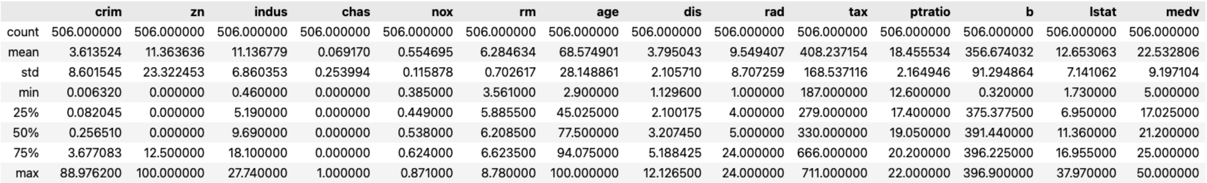
2 Programming Problems

2.1 Data Overview

In this section, I will show the data description. See Figure (1). There are thirteen attributes (independent variables, including one dummy) and one dependent variable *MEDV* in this dataset.



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

Figure 1: Data description

We want to figure out whether there are incomplete data in this dataset. I found that, all columns are complete in this dataset. So there's no need to process the Null (or NaN value). From the description of attributes, intuitively, I guess *TAX* is the most relevant attribute for *MEDV*. Since *TAX* represents the full-value property-tax rate, the higher the property value is, the higher tax will be in the housing market.

Then, we plot the *MEDV* distribution over each attribute, see Figure (5), in Appendix A.

From Figure (5), we observed that, there exist an apparent positive correlation between *MEDV* and *RM*, and a negative correlation between *MEDV* and *LSTAT*. Therefore, we can revise the guess in the previous paragraph correspondingly.

2.2 Data Processing

2.2.1 Pairwise Correlation Heatmap

Based on the given dataset, we draw a correlation heatmap, see Figure (6), in Appendix A.

We know that, it's highly likely to face the multicollinearity problem if the attributes are highly correlated. So we need to drop the attributes that have a high-pairwise correlation.

In this project, I choose **0.75** as the correlation threshold, which means, if an attribute has a pairwise correlation greater than 0.75, then it will be dropped from the attributes list. I dropped the following attributes: *indus*, *nox*, *dis*, *rad*, *tax*. After dropping by the correlation method, there are eight attributes left.

2.2.2 Data Scaling

The variables in the given dataset are measured in different scales, e.g., *chas* is a binary dummy, while *tax* measures the full-value property-tax rate per $10,000. The different scales lead to various contributions to the model, and hence, lead to measurement bias.

So, we use **sklearn.preprocessing.MinMaxScaler** to scale the data. See Figure 7 in Appendix A for the regression plot of *medv* over scaled and selected attributes, with 95% confidence interval.

2.3 Parameter Learning

We first split the data set into 80% training set, and 20$ testing set (validate set). In the training process, I use RMSE,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{N}}$$

as my loss function, where N is the number of observations.

And we use gradient descent as our learning method, we update the parameter $\mathbf{w}$ in each iteration. I will show the result in Section 2.4.

2.4 Results

For the first trial, here's my parameter setting.

- Step size α: 0.01
- Number of iterations *iters*: 1000
- Starting point $\mathbf{w}$: $[0, 0, \ldots, 0] \in \mathbb{R}^{1\times 9}$

Then, we have the following loss curve. See Figure 2

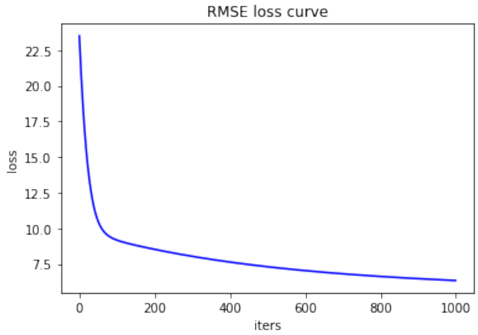


Figure 2: RMSE loss curve

We have the final $R^2 = 0.4789$, $RMSE = 0.4789$, and the estimated $\mathbf{w} = [9.72134189e+00, -1.75977438e+00, 5.64145608e+00, 3.37842047e+00, 1.18514503e+01, 8.10722372e-03, -2.70140031e+00, 1.03400465e+01, -6.57641608e+00]$.

As required, we repeat the above steps 10 times, changing the hyperparameters each time. Then, we have the loss curves in each trial. We want to figure out how the step size and the number of iterations will influence the $RMSE$.

See Figure 3 and Figure 4. Also, see Figure 8 in Appendix A for the $RMSE$ loss curves of ten independent trials with different hyperparameters.

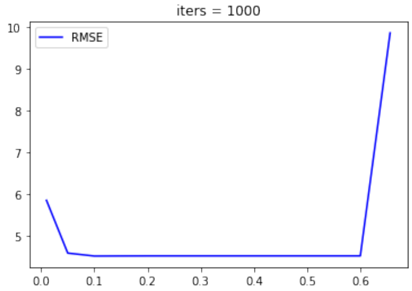


Figure 3: RMSE loss curve (fixed iters)

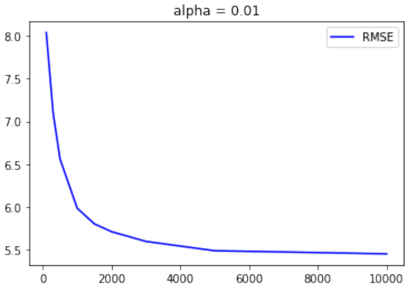


Figure 4: RMSE loss curve (fixed alpha)

We can have several remarks,

- When **fixing the number of iterations**, we will see two trends. When α is less than 0.05, there's a clear cutoff on $RMSE$ when α increases. This is because, the result does not reach the optimal within given iterations. However, when α is greater than 0.65, the result becomes less accurate.
- When **fixing** α, there's a clear downward change when increasing the iterations. This is because the small number of iterations can not lead to the optimal point. While the iterations are sufficient enough, increasing the iterations does not show an obvious change on $RMSE$.
- To **summarize**: increasing step size and decreasing the iterations will speed up the learning process, while the result becomes less accurate. Oppositely, decreasing the step size and increasing the iterations will definitely make the model run slower, but will have a more accurate result.

Appendices

A Figures

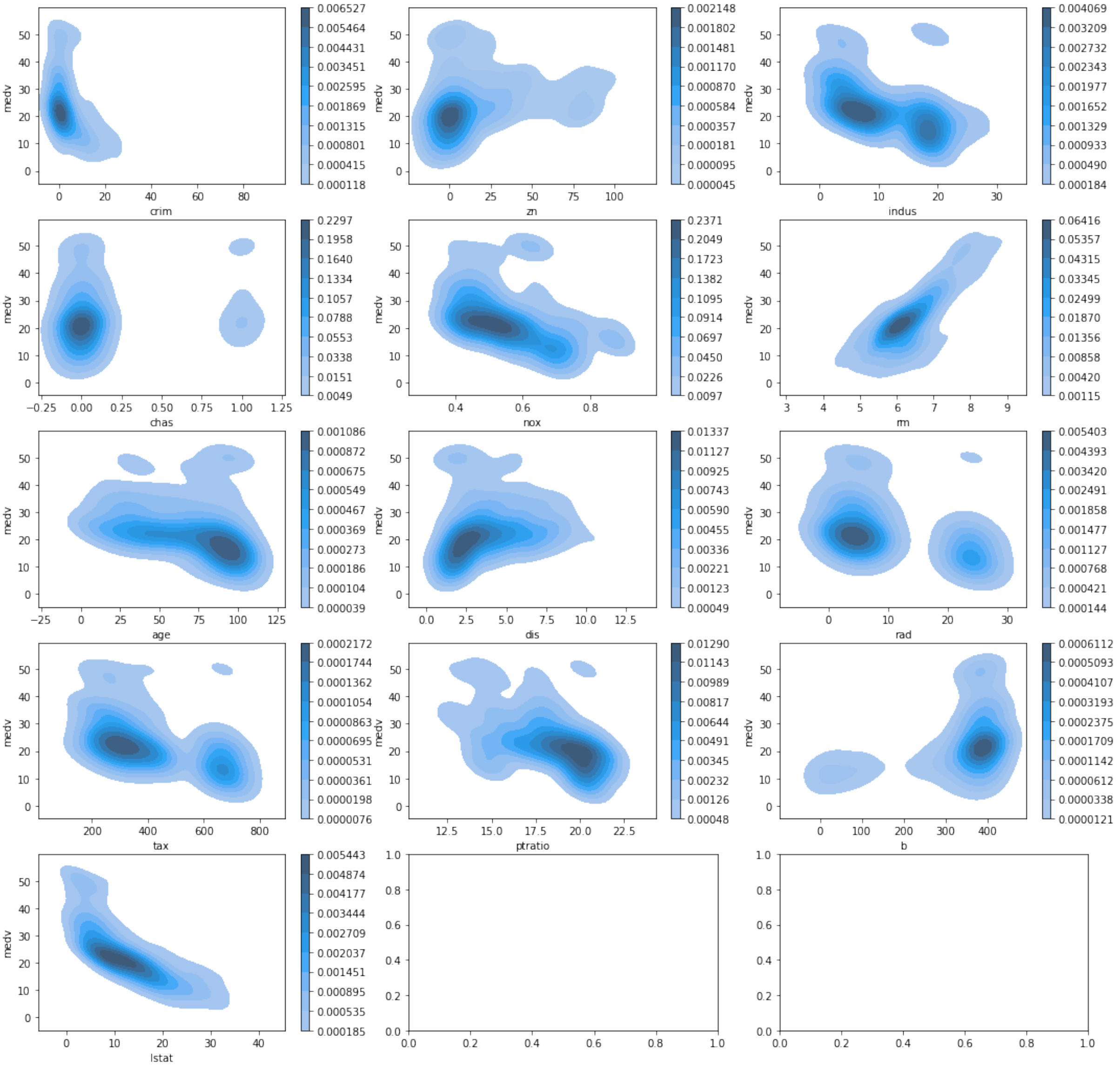
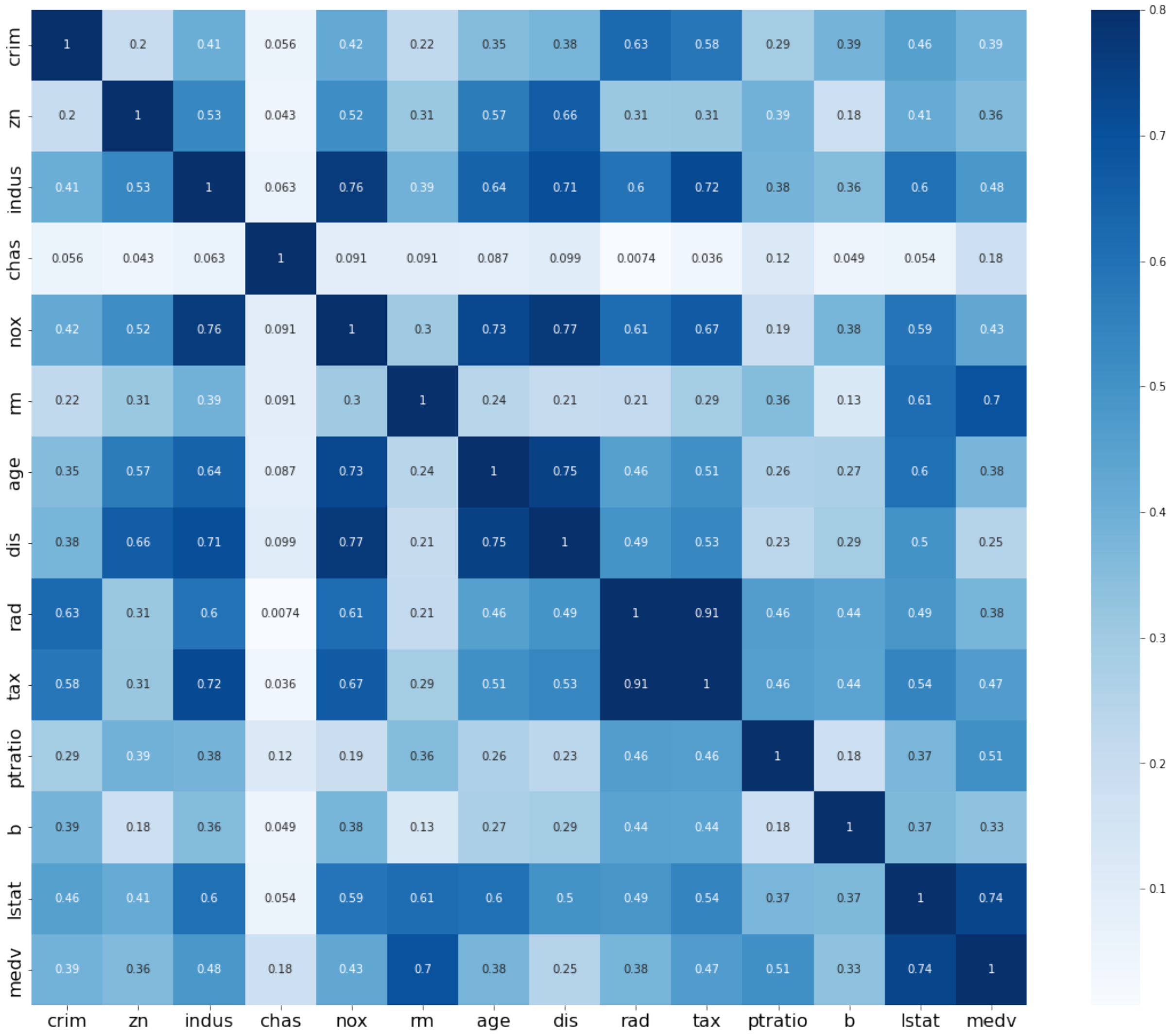


Figure 5: MEDV distribution over each attribute



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
crim	1	0.2	0.41	0.056	0.42	0.22	0.35	0.38	0.63	0.58	0.29	0.39	0.46	0.39
zn	0.2	1	0.53	0.043	0.52	0.31	0.57	0.66	0.31	0.31	0.39	0.18	0.41	0.36
indus	0.41	0.53	1	0.063	0.76	0.39	0.64	0.71	0.6	0.72	0.38	0.36	0.6	0.48
chas	0.056	0.043	0.063	1	0.091	0.091	0.087	0.099	0.0074	0.036	0.12	0.049	0.054	0.18
nox	0.42	0.52	0.76	0.091	1	0.3	0.73	0.77	0.61	0.67	0.19	0.38	0.59	0.43
rm	0.22	0.31	0.39	0.091	0.3	1	0.24	0.21	0.21	0.29	0.36	0.13	0.61	0.7
age	0.35	0.57	0.64	0.087	0.73	0.24	1	0.75	0.46	0.51	0.26	0.27	0.6	0.38
dis	0.38	0.66	0.71	0.099	0.77	0.21	0.75	1	0.49	0.53	0.23	0.29	0.5	0.25
rad	0.63	0.31	0.6	0.0074	0.61	0.21	0.46	0.49	1	0.91	0.46	0.44	0.49	0.38
tax	0.58	0.31	0.72	0.036	0.67	0.29	0.51	0.53	0.91	1	0.46	0.44	0.54	0.47
ptratio	0.29	0.39	0.38	0.12	0.19	0.36	0.26	0.23	0.46	0.46	1	0.18	0.37	0.51
b	0.39	0.18	0.36	0.049	0.38	0.13	0.27	0.29	0.44	0.44	0.18	1	0.37	0.33
lstat	0.46	0.41	0.6	0.054	0.59	0.61	0.6	0.5	0.49	0.54	0.37	0.37	1	0.74
medv	0.39	0.36	0.48	0.18	0.43	0.7	0.38	0.25	0.38	0.47	0.51	0.33	0.74	1

Figure 6: Correlation heatmap

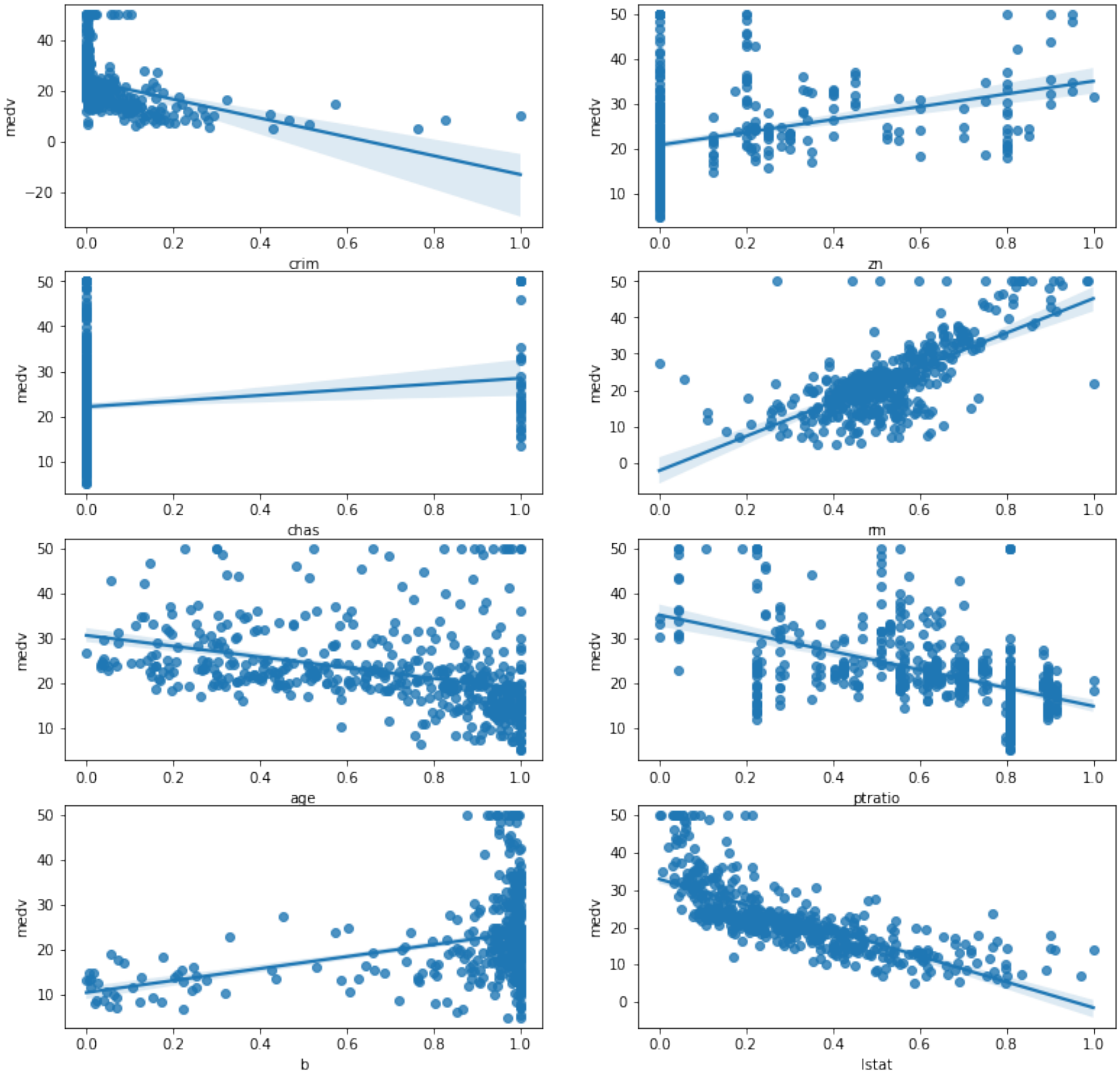
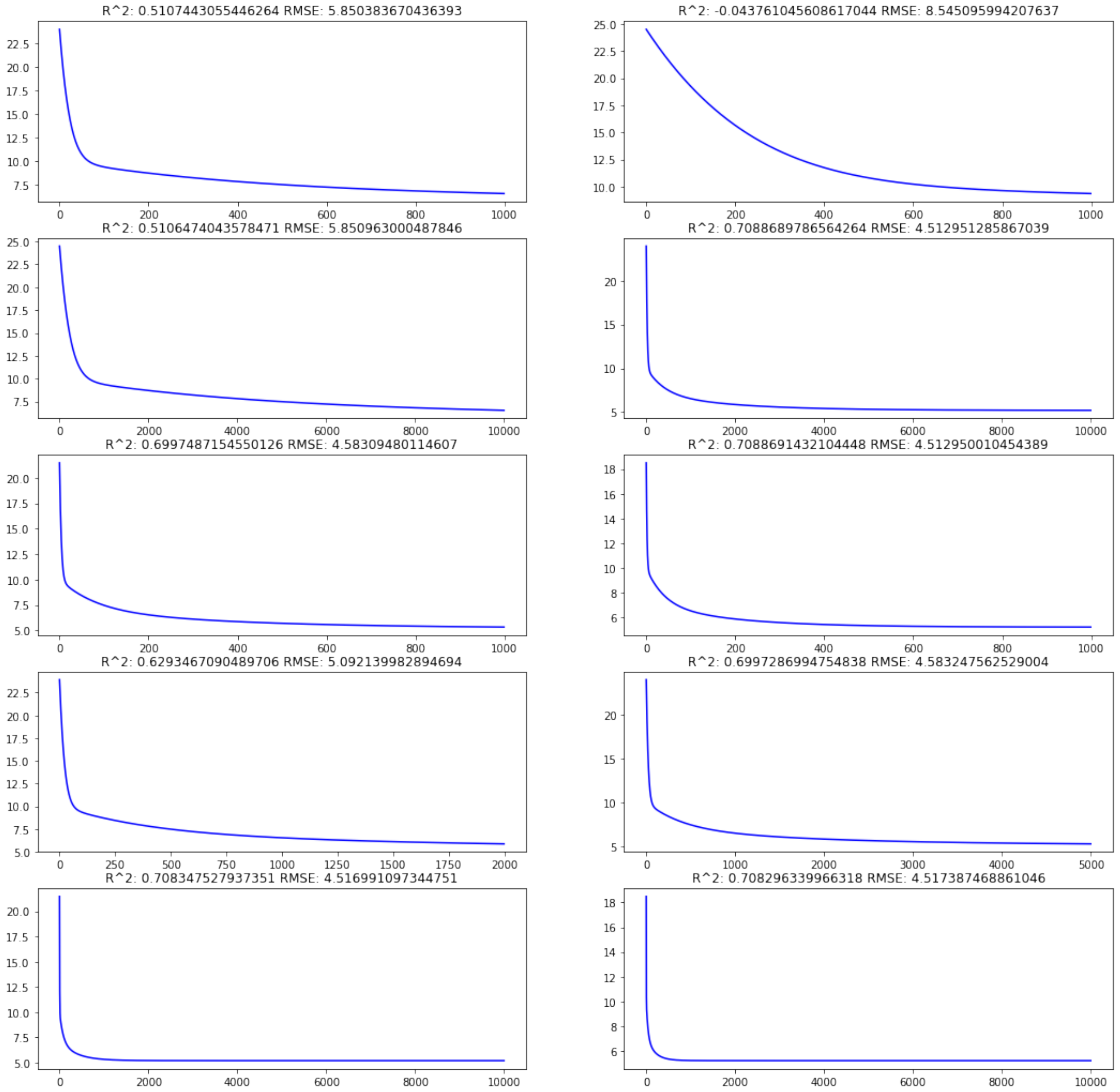


Figure 7: Regression plots



R^2: 0.5107443055446264 RMSE: 5.850383670436393

R^2: -0.043761045608617044 RMSE: 8.545095994207637

R^2: 0.5106474043578471 RMSE: 5.850963000487846

R^2: 0.7088689786564264 RMSE: 4.512951285867039

R^2: 0.6997487154550126 RMSE: 4.58309480114607

R^2: 0.7088691432104448 RMSE: 4.512950010454389

R^2: 0.6293467090489706 RMSE: 5.092139982894694

R^2: 0.6997286994754838 RMSE: 4.583247562529004

R^2: 0.708347527937351 RMSE: 4.516991097344751

R^2: 0.708296339966318 RMSE: 4.517387468861046

Figure 8: Ten trials